NeuroImage: Clinical 9 (2015) 640–647

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Quantifying brain tissue volume in multiple sclerosis with automated lesion segmentation and filling

Sergi Valverde [a,*], Arnau Oliver [a], Eloy Roura [a], Deborah Pareto [b], Joan C. Vilanova [c], Lluís Ramió-Torrentà [d], Jaume Sastre-Garriga [e], Xavier Montalban [e], Àlex Rovira [b], Xavier Lladó [a]

[a] Dept. of Computer Architecture and Technology, University of Girona, Spain
[b] Magnetic Resonance Unit, Dept. of Radiology, Vall d'Hebron University Hospital, Spain Architecture and Technology, University of Girona, Spain
[c] Girona Magnetic Resonance Center, Spain
[d] Multiple Sclerosis and Neuro-immunology Unit, Dr. Josep Trueta University Hospital, Spain
[e] Neurology Unit, Multiple Sclerosis Centre of Catalonia (Cemcat), Vall d'Hebron University Hospital, Spain

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 21 October 2015
Accepted 23 October 2015
Available online 28 October 2015

Keywords:
Brain
Multiple sclerosis
MRI
Brain atrophy
Automated tissue segmentation
White matter lesions
Lesion filling

ABSTRACT

Lesion filling has been successfully applied to reduce the effect of hypo-intense T1-w Multiple Sclerosis (MS) lesions on automatic brain tissue segmentation. However, a study of fully automated pipelines incorporating lesion segmentation and lesion filling on tissue volume analysis has not yet been performed. Here, we analyzed the % of error introduced by automating the lesion segmentation and filling processes in the tissue segmentation of 70 clinically isolated syndrome patient images. First of all, images were processed using the LST and SLS toolkits with different pipeline combinations that differed in either automated or manual lesion segmentation, and lesion filling or masking out lesions. Then, images processed following each of the pipelines were segmented into gray matter (GM) and white matter (WM) using SPM8, and compared with the same images where expert lesion annotations were filled before segmentation. Our results showed that fully automated lesion segmentation and filling pipelines reduced significantly the % of error in GM and WM volume on images of MS patients, and performed similarly to the images where expert lesion annotations were masked before segmentation. In all the pipelines, the amount of misclassified lesion voxels was the main cause in the observed error in GM and WM volume. However, the % of error was significantly lower when automatically estimated lesions were filled and not masked before segmentation. These results are relevant and suggest that LST and SLS toolboxes allow the performance of accurate brain tissue volume measurements without any kind of manual intervention, which can be convenient not only in terms of time and economic costs, but also to avoid the inherent intra/inter variability between manual annotations.

1. Introduction

Multiple sclerosis (MS) is associated with irreversible brain damage not only in demyelinated plaques, but also in normal-appearing gray matter (GM) and white matter (WM), where recent studies have shown that the rate of tissue loss per year in MS patients ranges from 0.7% to 1.6% in GM, and 0.6% to 0.9% in WM (Filippi et al., 2013; Pérez-Miralles et al., 2013; Sastre-Garriga et al., 2014). Given the correlation between brain atrophy and disease disability, measuring the change in tissue volume is clinically relevant because it allows for optimizing possible treatments and patient management in early stages of the disease (Filippi et al., 2013; Sastre-Garriga et al., 2014; Uher et al., 2014).

Automated tissue segmentation techniques based on magnetic resonance imaging (MRI) such as the Statistical Parametric Mapping (SPM) (Ashburner and Friston, 2005), FAST (Zhang et al., 2001), or SIENA-X (Smith et al., 2002) are currently standard tools to assess brain tissue volume (De Bresser et al., 2011; Valverde et al., 2015a). The reproducibility of these techniques has been analyzed in several studies using scan–rescan measurement tests, reporting mean percentages of error in FAST GM of −0.22% (De Boer et al., 2010), 0.05% (De Boer et al., 2010) and −0.80% (Nakamura et al., 2014) in SPM8 GM, 1.50% (Nakamura et al., 2014) in SIENA-X GM, 0.13% (De Boer et al., 2010) in FAST WM, and 0.25% (De Boer et al., 2010) in SPM WM. However, existing differences for a particular method in different studies may be influenced by the same image data, imaging hardware and acquisition parameters (Clark et al., 2006). Furthermore, several authors have reported that the inclusion of WM lesions in tissue segmentation can affect significantly the accuracy of these techniques (Battaglini et al., 2012; Nakamura and Fisher, 2009; Valverde et al., 2015b), leading to

* Corresponding author at: Ed. P-IV, Campus Montilivi, University of Girona, 17071 Girona, Spain.
E-mail address: svalverde@eia.udg.edu (S. Valverde).

http://dx.doi.org/10.1016/j.nicl.2015.10.012

the development of different preprocessing strategies to fill lesion regions with signal intensities similar to WM before tissue segmentation (Battaglini et al., 2012; Chard et al., 2010; Valverde et al., 2014). So far, in all the lesion filling approaches, MS lesions have to be delineated first, usually by their manual annotation, which is a tedious, challenging and time-consuming task (Sanfilipo et al., 2005). This fact and the necessity to analyze quantitatively focal MS lesions in individual (Cabezas et al., 2014) and temporal (Ganiler et al., 2014) studies have been driving in recent years the development of automated new lesion segmentation techniques (García-Lorenzo et al., 2013; Guizard et al., 2015; Lladó et al., 2012).

Although lesion filling techniques have already been applied to assess the progression of GM atrophy of MS patients (Ceccarelli et al., 2012; Nakamura et al., 2014; Popescu et al., 2014), still an extensive analysis of the effect of fully automated pipelines, incorporating both automated MS lesion segmentation and posterior lesion filling on tissue segmentation methods has not yet been performed. In this study, we analyze the effect of two publicly available automated pipelines, Salem Lesion Segmentation (SLS) (Roura et al., 2015) and Lesion Segmentation Toolbox (LST) (Schmidt et al., 2012), on the accuracy of the GM and WM volume estimations of a cohort of 70 clinically isolated syndrome (CIS) patients. For each automated pipeline, we evaluate the deviation in GM and WM volume between images where manual expert annotations have been used to refill lesions before tissue segmentation with SPM8 (Ashburner and Friston, 2005), and the same images where lesions have been automatically segmented and either masked or lesion filled before tissue segmentation.

2. Materials and methods

2.1. Image acquisition

Seventy CIS patients from the same center (Hospital Vall D'Hebron, Barcelona (Spain)) in which the clinical presentation was clearly suggestive of multiple sclerosis underwent MR imaging on the same 3 T Siemens with 12-channel phased-array head coil (Trio Tim, Siemens, Germany). The following pulse sequences were obtained: 1) transverse proton density and T2-weighted fast spin-echo (TR = 2500 ms, TE = 16–91 ms, voxel size = $0.78 \times 0.78 \times 3$ mm^3); 2) transverse fast T2-FLAIR (TR = 9000 ms, TE = 93 ms, TI = 2500 ms, flip angle = 120°, voxel size = $0.49 \times 0.49 \times 3$ mm^3); and 3) sagittal 3D T1 magnetization prepared rapid gradient-echo (MPRAGE) (TR = 2300 ms, TE = 2 ms; flip angle = 9°; voxel size = $1 \times 1 \times 1.2$ mm^3). White matter lesion masks were semi-automatically delineated from either PD-w (46 patients) or FLAIR (24 patients) images using JIM software (Xinapse Systems, http://www.xinapse.com/home.php) by an expert radiologist of the same hospital center with more than 10 years of experience. Mean lesion volume was 4.1 ± 4.7 ml (range 0.2–18.3 ml), and 3.65 ± 3.94 ml (range 0.1–18.3 ml) on PD-w and FLAIR images, respectively.

2.2. Automated lesion segmentation and filling

Automated lesion segmentation and filling was performed using the T1-w and FLAIR image modalities on two publicly available toolkits implemented for the SPM (http://www.fil.ion.ucl.ac.uk/spm) software package:

2.2.1. SLS toolbox

The SLS pipeline (http://atc.udg.edu/salem/slsToolbox/index.html) was composed of the following automated steps: T1-w and FLAIR images were first skull-stripped and intensity corrected using the Brain Extraction Tool (BET) (Smith, 2002) with optimized parameter choice as described in Popescu et al. (2012), and the N3 method (Sled et al., 1998), respectively. Corrected T1-w and FLAIR images were then linearly co-registered (12-parameter affine) using internal SPM routines, with normalized mutual information as objective function and trilinear interpolation with no wrapping. Lesion segmentation was performed by an initial tissue segmentation of the T1-w image to separate lesions from tissue, followed by a thresholding step and a regionwise refinement of the FLAIR image (Roura et al., 2015). The initial parameter used to adjust the detected candidate lesions was set to $\alpha = 2$, while the percentage of lesion candidate regions to belong to WM and GM over cerebro spinal fluid (CSF), percentage of neighbor voxels belonging to WM, and candidate size was set to $\lambda\_ts = 0.7$, $\lambda\_nb = 0.6$, and size = 3 mm^3. Estimated lesion masks were then automatically filled using the method (Valverde et al., 2014), where candidate region voxels were replaced by random values of a normal distribution generated from the mean normal-appearing WM signal intensity of each two-dimensional T1-w slice. The SLF method was run with default parameters.

2.2.2. LST toolbox

The LST pipeline (www.applied-statistics.de/lst) was composed of the following automated steps: T1-w and FLAIR images were skull-stripped and intensity-corrected using the VBM8 toolbox included also as part of the SPM package. Afterwards, corrected T1-w and FLAIR images were linearly (12-parameter affine) and non-linearly co-registered using also internal SPM8 routines. Lesion segmentation was performed by computing an initial tissue segmentation of the T1-w image to compute a lesion belief map based on the FLAIR and T1-w images (Schmidt et al., 2012). This map was refined iteratively weighting the likelihood of belonging to WM or GM against the likelihood of belonging to lesions until no further voxels were assigned to lesions. The required initial threshold kappa was set to $k = 0.15$, while the lesion belief map was set to lbm = GM. Estimated lesion masks were then automatically filled using an internal filling method inspired by a previous technique proposed in Chard et al. (2010), where candidate region voxels where replaced by random intensities from a Gaussian distribution generated from the normal-appearing WM intensities and then filtered to reintroduce the original spatial variation in WM.

2.3. Tissue volume analysis

All images were processed with both toolboxes and compared independently in order to preserve the differences in the internal routines of each toolbox. First, T1-w images processed by the SLS toolbox (see Table 1(a)) were segmented into GM, WM and CSF volumes using SPM8 after following five different pipeline configurations that differed in the level of manual intervention: 1) Original images were segmented including WM lesions (*Original* pipeline); 2) Expert manual lesion annotations were masked before tissue segmentation and relabeled as WM after (*Expert masked* pipeline); 3) Estimated lesion masks provided by the SLS method were masked before tissue segmentation and relabeled as WM after (*SLS masked* pipeline); 4) Estimated lesion masks provided by the SLS method were filled with the SLF method before tissue segmentation (*SLS filled*); and 5) Expert manual lesion annotations were filled before tissue segmentation and used as ground-truth images (*Expert filled* pipeline). In the case of the pipelines where lesions voxels were masked, either with automatic or manual annotations, lesion masks were used to remove lesion voxels in the T1-w image. Therefore, those voxels were not considered during tissue segmentation and were added to the WM class after it to maintain the actual brain volume of each patient. In contrast, in the lesion filling pipelines, automatic or manual lesion annotations were used to refill the correspondent T1-w image voxels with signal intensities similar to the WM, and lesion voxels were considered as normal-appearing WM in tissue segmentation.

All resultant tissue probability maps were binarized into GM, WM and CSF masks by extracting the maximum probability for each particular tissue. GM and WM tissue volume was computed by multiplying the number of voxels in binary masks by the voxel size ($1 \times 1 \times 1.2$ mm^3). Volume measures were normalized to correct the differences between subjects by dividing the GM and WM volume by

Table 1
Evaluation pipelines followed in the present study. The set of T1-w images is processed independently for either the SLS (a) and LST (b) toolboxes. First, T1-w images are preprocessed (skull stripped and intensity corrected) using the routines indicated by each toolbox. Then, the preprocessed images are segmented into CSF, GM and WM tissue using SPM8 after following five different pipelines that differ in the level of manual intervention: 1) images are segmented including WM lesions (*Original* pipeline), 2) Expert manual lesion annotations are masked before tissue segmentation (*Expert masked* pipeline), 3) Estimated lesion masks returned by the same toolbox are masked before tissue segmentation (*SLS/LST masked* pipeline), 4) Estimated lesion masks returned by the same toolbox are filled with the lesion-filling method incorporated by each pipeline (*SLS/LST filled* pipeline), and 5) Expert manual lesion annotations are filled before tissue segmentation and used as ground-truth images (*Expert filled* pipeline).

(a)

Pipeline	Preprocessing	Lesion segmentation	Lesion filling	Tissue segmentation
1. *Original*	BET + N3	–	–	SPM8
2. *Expert masked*	BET + N3	Manual	Expert annotations are masked	SPM8
3. *SLS masked*	BET + N3	SLS	SLS lesion masks are masked	SPM8
4. *SLS filled*	BET + N3	SLS	SLS lesion masks are filled by SLF	SPM8
5. *Expert filled* (GT)	BET + N3	Manual	Expert annotations are filled by SLF	SPM8

(b)

Pipeline	Preprocessing	Lesion segmentation	Lesion filling	Tissue segmentation
1. *Original*	SPM8	–	–	SPM8
2. *Expert masked*	SPM8	Manual	Expert annotations are masked	SPM8
3. *LST masked*	SPM8	LST	LST lesion masks are masked	SPM8
4. *LST filled*	SPM8	LST	LST lesion masks are filled by LST	SPM8
5. *Expert filled* (GT)	SPM8	Manual	Expert annotations are filled by LST	SPM8

the whole brain volume. Then, the percent (%) absolute error in total and normal-appearing GM and WM volume was computed between pipelines: *Original* versus *Expert filled* images, *Expert masked* versus *Expert filled*, *SLS masked* versus *Expert filled*, and *SLS filled* versus *Expert filled*. The absolute error in total and normal-appearing GM and WM volume for each automated pipeline were computed using the following equations:

$$GM_{\{1\ldots4\}vs5} = \frac{|NGMV_{\{1\ldots4\}} - NGMV_5|}{NGMV_5} \times 100$$

$$WM_{\{1\ldots4\}vs5} = \frac{|NWMV_{\{1\ldots4\}} - NWMV_5|}{NWMV_5} \times 100$$

where $NGMV_{\{1\ldots4\}}$ and $NWMV_{\{1\ldots4\}}$ refer to the normalized GM and WM tissue volume, and the sub-indexes indicate the pipeline used: (1) *Original*, (2) *Expert masked*, (3) *SLS masked*, (4) *SLS filled and* (5) *Expert filled* pipeline used as ground-truth. Normal-appearing GM and WM volume was computed similarly, but lesion voxels were not considered in normalized GM and WM volume estimations. The procedure was then repeated identically for the LST toolbox (see Table 1(b)).

2.4. Statistical analysis

Statistical analysis was performed using the Matlab software package (http://es.mathworks.com/products/matlab). Differences in GM and WM volume of each evaluated pipeline were analyzed using a repeated measures ANOVA model with 3 degrees of freedom for the time variable and 207° for the error, followed by a series of post-hoc pairwise significant t-tests with Bonferroni correction between methods. Moreover, the Pearson's linear correlation coefficient was used to compute the correlation between % differences in GM and WM and lesion volume, and between % differences in GM and WM and the error produced by the automated lesion segmentation methods (Error I type: number of false positive outcomes, and Error II type: false negative outcomes). In all the analysis, we considered data significant at p-values < 0.05.

3. Results

3.1. Differences in tissue volume

First, we analyzed the differences in total tissue volume between the images processed following each of the SLS pipelines and the images where expert lesion masks had been filled with the SLF method before tissue segmentation. Automated lesion segmentation and filling reduced significantly the % of error in total GM ($p < 0.032$) on the images processed with the fully automated *SLS filled* pipeline when compared with the same images segmented including lesions (*Original* pipeline) (see Fig. 1A). Similarly, the % differences in total WM were also significantly lower on the *Expert masked* ($p < 0.040$) and *SLS filled* ($p < 0.002$) pipelines when compared with the *Original* images (see Fig. 1B). Differences in total GM and WM between the *SLS masked* and *SLS filled* pipelines were not statistically different.

Regarding the LST toolbox, the mean % of error in GM volume was <0.12% in all the evaluated pipelines and similar to the values reported previously by the SLS, but was significantly higher in the *Original* images ($p < 0.003$) (see Fig. 1C). In WM, the effect of hypo-intense lesions was also significantly higher in the *Original* images ($p < 0.001$) when compared with the rest of the pipelines (see Fig. 1D). As in the SLS, the differences in total GM and WM between *LST masked* and *LST filled* were not significant.

The observed % of error in total GM and WM volume was not only distributed in lesion regions but also in normal-appearing tissue (see Fig. 2). In all the evaluated pipelines but the *Expert masked*, normal-appearing WM was overestimated by the effect of hypo-intense lesion voxels that were still present before tissue segmentation, either because they were not processed intentionally (*Original* pipeline), or as the result of misclassified lesion voxels. Lesion voxels that were classified as WM shifted down the signal intensity threshold between GM and WM and caused the actual GM voxels presenting an intensity profile similar to that of the lesions to be reassigned to WM. Identically; normal-appearing GM was underestimated by the opposite effect of lesion voxels in GM tissue volume. More importantly, in the images processed with the *Original*, *SLS masked*, and *SLS filled* pipelines, the actual % of error in total GM and WM volume was partly canceled between the opposite directions of the errors produced in normal-appearing tissue and the number of remaining lesion voxels that were incorrectly classified as GM (see Fig. 2).

As expected, images where expert lesion masks were masked before segmentation (*Expert masked* pipeline) returned the lowest % of error in normal-appearing GM (see Fig. 2A) and WM (see Fig. 2B) when compared not only with *Original* images ($p < 0.001$), but also with images processed with the *SLS masked* pipeline ($p < 0.018$). The % differences in normal-appearing WM of the images where estimated lesions using SLS were filled were significantly lower than in the same images where lesions were masked ($p < 0.024$). In contrast, differences were

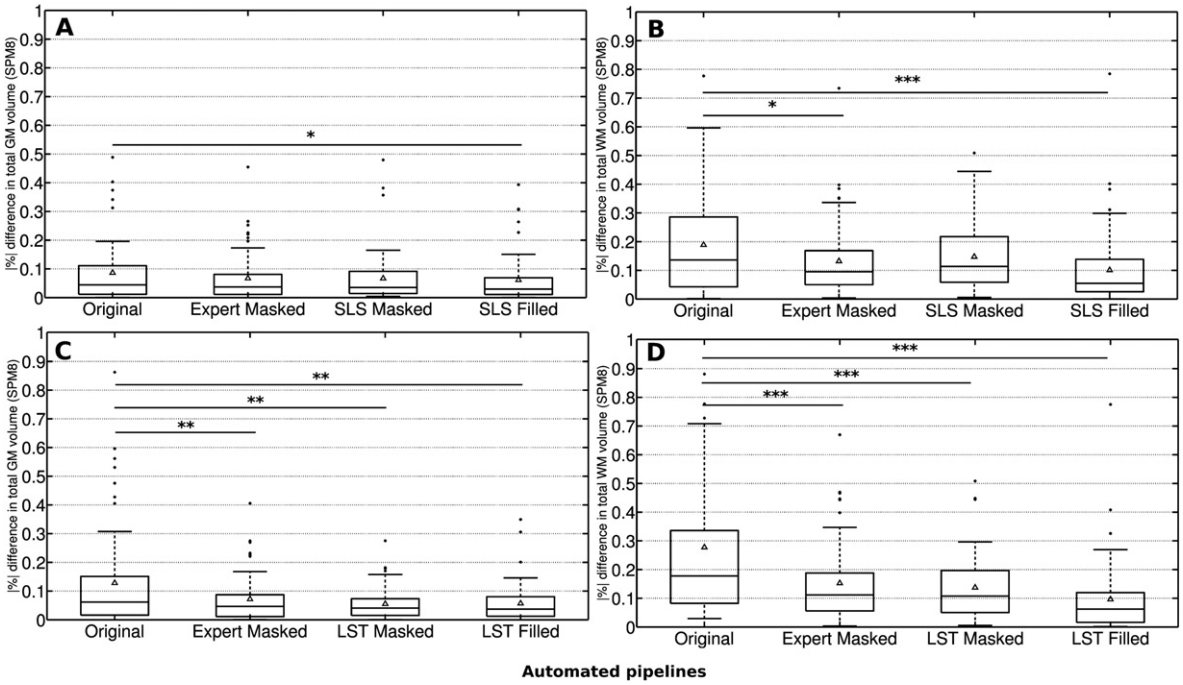

Fig. 1. % of absolute error in total GM and WM volume between segmented images where the annotated lesion masks were refilled before tissue segmentation (*Expert filled*) and the same images processed following the *Original*, *Expert masked*, *SLS/LST masked*, and *SLS/LST filled* pipelines. Results for the SLS toolbox are shown in the top row for GM (A) and WM (B), and for the LST toolbox in the bottom row for GM (C) and WM (D). The Δ symbol depicts the mean % difference in total GM/WM tissue for each pipeline. Horizontal lines show significant differences between evaluated pipelines with *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

similar for both tissues between the fully automated *SLS filled* and the *Expert masked* pipelines, showing that refilled voxels reduced the effect of hypo-intense lesions in normal-appearing tissue.

Similarly, in LST pipelines part of the % differences in total GM and WM was also partly canceled by the opposite direction of the errors in normal-appearing and the remaining lesion voxels that were incorrectly classified as GM. As expected, the % differences in normal-appearing GM (see Fig. 2C) and WM (see Fig. 2D) were lower in the *Expert masked* pipeline ($p < 0.024$), due to the null effect of hypo-intense lesions in tissue segmentation. As in SLS, the effect of masking expert lesion masks on the errors in tissue segmentation was similar to that in the automated lesion segmentation and filling. The % differences in

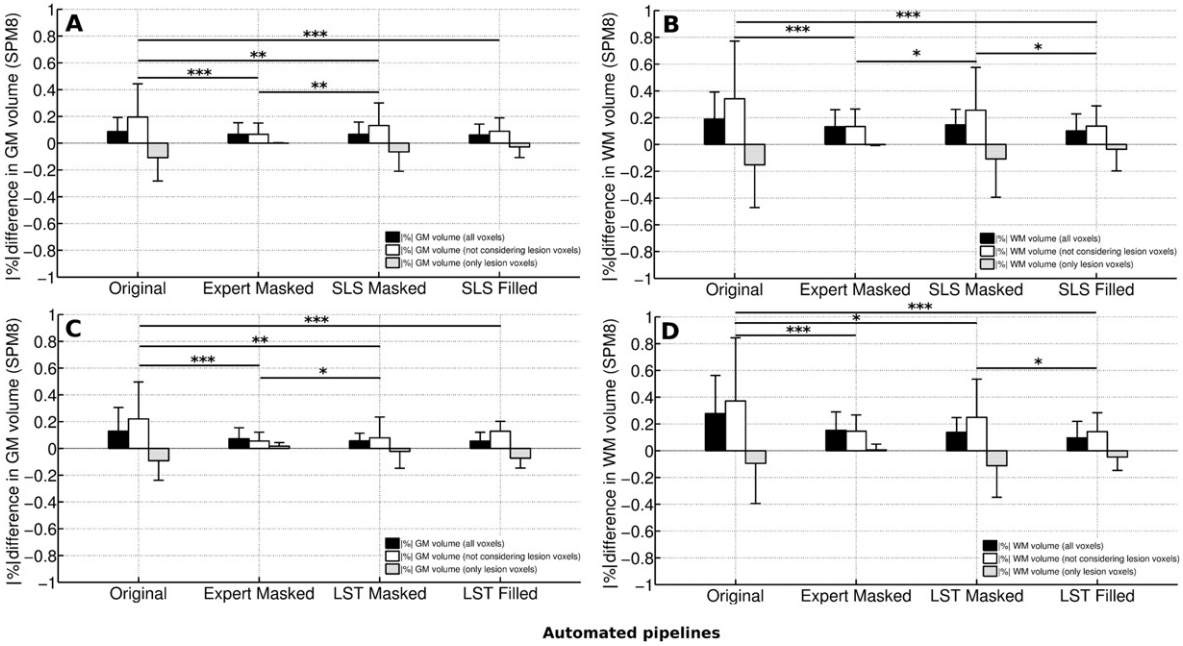

Fig. 2. Mean % of absolute error in tissue volume between segmented images where the annotated lesion masks were refilled before tissue segmentation (*Expert filled*) and the same images processed following each of the evaluated pipelines. Results for the SLS toolbox are shown in the top row for GM (A) and WM (B), and for the LST toolbox in the bottom row for GM (C) and WM (D). Differences in tissue volume are split in three regions: those produced when all voxels are considered (black bars), those produced when not considering lesion voxels (white bars), and those produced only in lesion voxels (gray bars). Lesion regions bars are plotted with negative bars to visualize the opposite direction of the errors in lesion voxels with respect to normal-appearing tissue. Vertical lines at each bar depict the % standard deviation difference in tissue volume for each pipeline. Horizontal lines show significant differences in normal-appearance volume between evaluated pipelines with *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

normal-appearing WM of the images where estimated lesions using LST were filled were also significantly lower than in the same images where lesions were masked ($p < 0.048$).

Fig. 3 depicts for a single patient image, the differences in the overlap of the tissue segmentation classes for each evaluated pipeline and the pertinent *Expert filled* image used as a ground-truth. As expected, the error in normal-appearing tissue (shown in red) was the lowest in the masked pipelines (images F and J), while the number of misclassified lesion voxels (shown in green) was remarkably higher in the *Original* pipelines (images E and I). This fact showed that the inclusion of hypo-intense lesion voxels into the tissue distributions has a clear effect in the misclassification of the normal-appearing tissue between boundaries, and also produces changes in the segmentation of brain structures such as the putamen. In contrast, when compared to these pipelines, the number of misclassified voxels in the automated pipelines incorporating lesion filling (panels H and L) was remarkably lower, although some false negatives were still present in the segmentation due to errors in the automatic lesion segmentation. The number of misclassified voxels was moderately lower in the automated pipelines incorporating lesion filling, when compared with automated pipelines where lesion masks were masked before segmentation (images G an K), although those differences were hardly appreciated in the picture.

When analyzing the % differences in tissue volume between LST and SLS pipelines, we observed that differences in GM between the evaluated pipelines were not significant. In contrast, the % differences between *masked* and *filled* pipelines were found significant for total WM between *LST filled* and *SLS masked* ($p < 0.191$), normal-appearing WM between *LST masked* and *SLS filled* ($p = 0.007$), and normal-appearing WM between *Lst filled* and *SLS masked* ($p < 0.002$).

Finally, we studied the effect of the image modality used to annotate the expert lesion masks in the overall result. We recomputed the differences in total and normal-appearing GM and WM volume for the two subsets of images where expert masks were annotated using PD-w or FLAIR images. The differences in GM and WM volume between subsets were not statistically different for any of the SLS or LST evaluated pipelines ($p > 0.42$).

3.2. Correlation with lesion volume

We also analyzed the extent to which lesion volume affected the normal-appearing GM and WM volume measurements of each of the evaluated pipelines. Lesion volume strongly correlated with the reported % of error in GM and WM in the *Original*, *Expert masked* and *SLS filled* pipelines ($r > 0.77$, $p < 0.001$), and moderately in the *SLS masked* ($r > 0.41$, $p < 0.001$). However, the effect of lesion volume was different for each evaluated pipeline (Fig. 4A).

As expected, the deviation in normal-appearing GM and WM volume was remarkably higher in the images segmented with lesions,

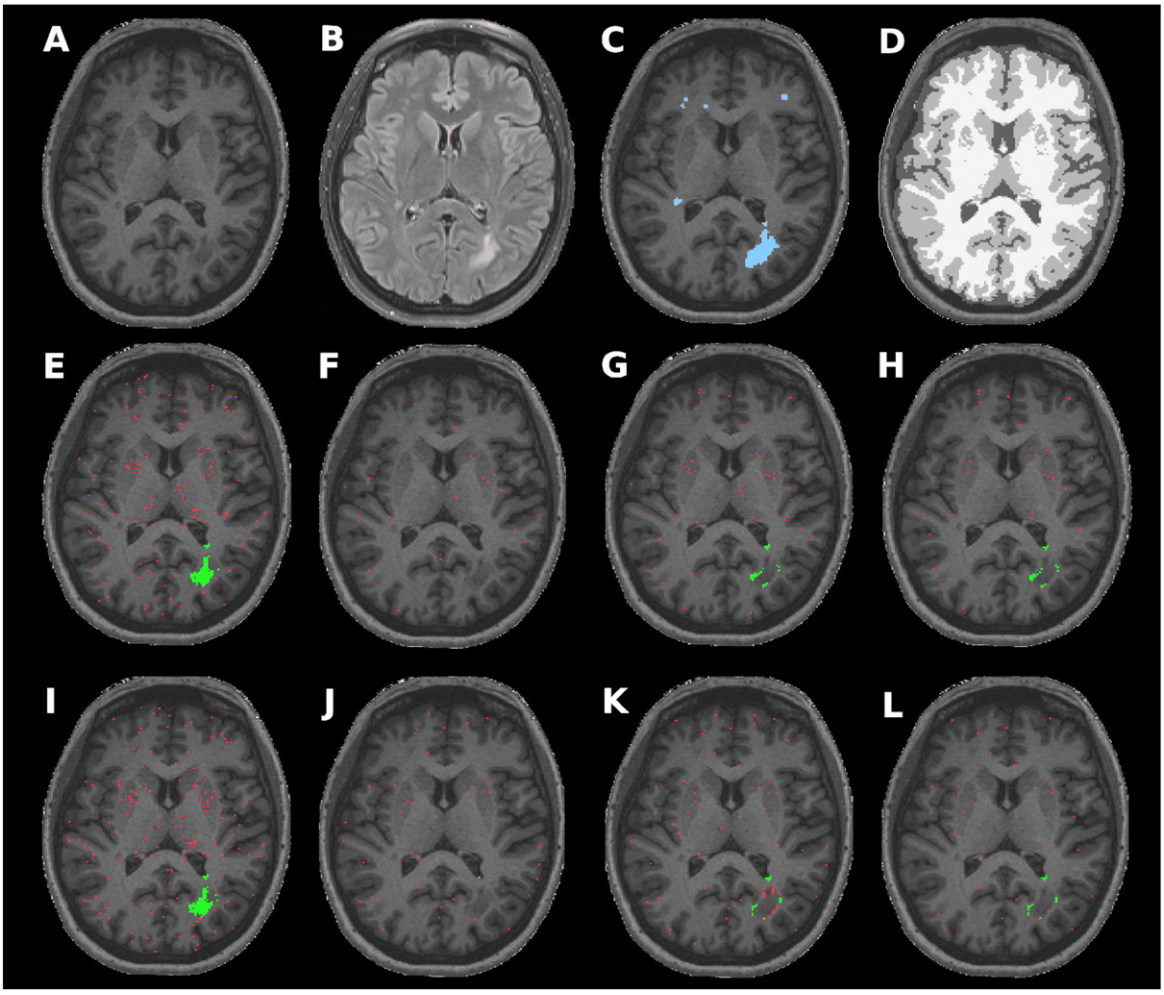

Fig. 3. For a single patient image of the dataset, we show the differences in the overlap of the tissue segmentation classes for each evaluated pipeline and the pertinent *Expert filled* image used as a ground-truth. Differences for any tissue class with respect to *Expert filled* are represented in green for lesion voxels, and in red for normal-appearing voxels. First row: input T1-w (A), input FLAIR (B), T1-w with expert annotations highlighted in blue (C), and T1-w output segmentation for the *Expert filled* image with CSF, GM and WM voxels depicted in black, gray, and white, respectively. Second row: for the images processed with the SLS toolkit, differences in any tissue classes for the *Original* (E), *Masked* (F), *SLS masked* (G), and *SLS filled* (H) pipelines. Third row: for the images processed with the LST toolkit, differences in any tissue class for the *Original* (I), *Masked* (J), *LST masked* (K), and *LST filled* (L) pipelines.

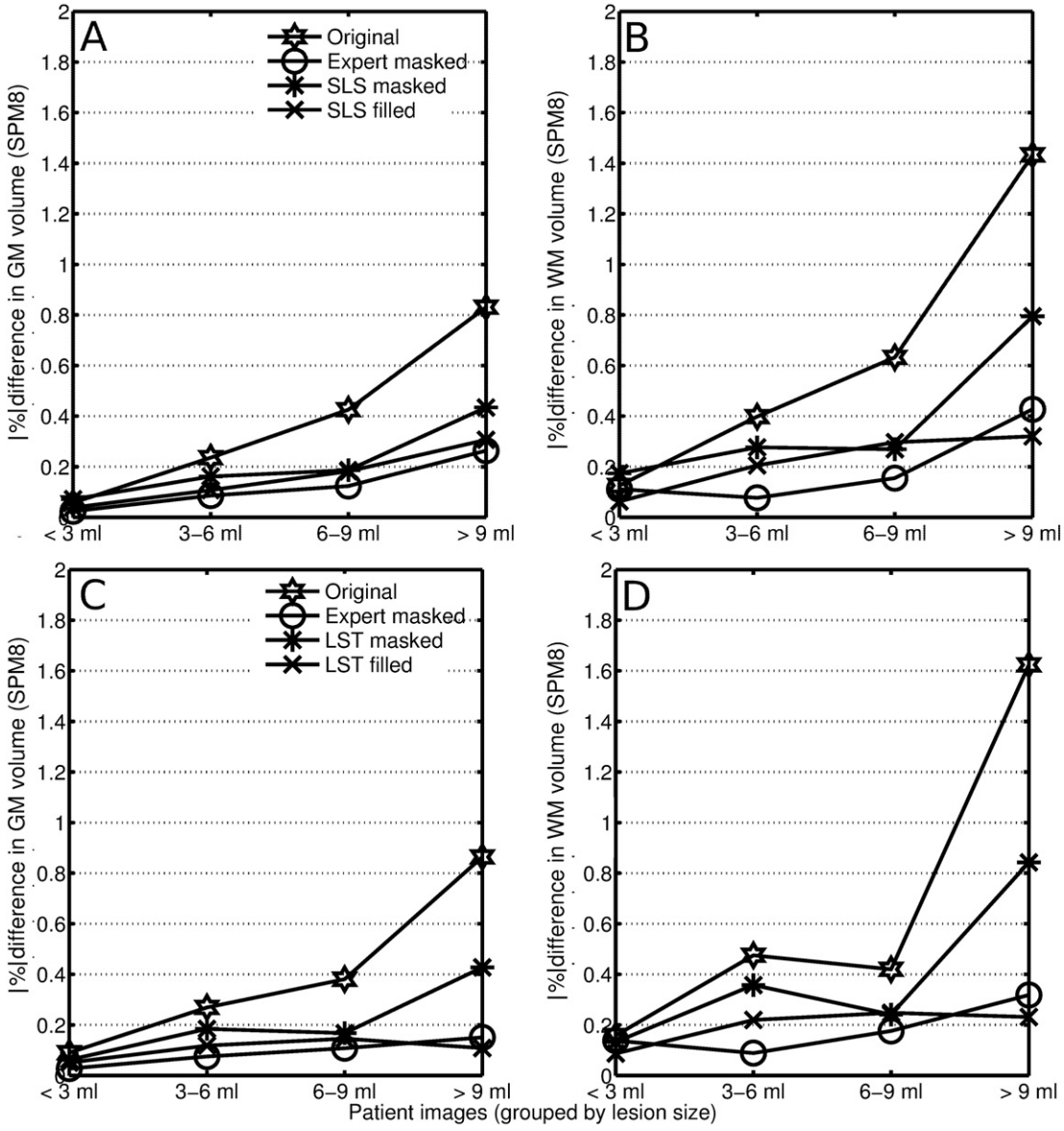

Fig. 4. Mean % of absolute error in normal-appearing GM and WM volume split by image groups with lesion size in the range (<3 ml, 3–6 ml, 6–9 ml, and >9 ml). Values for each group represent the mean % error between the images processed with the *Expert filled* and each of the evaluated pipelines (*Original* (✡), *Expert masked* (○), *SLS/LST masked* (✱), and *SLS/LST filled* (✖)). Results for the SLS toolbox are shown in the top row for GM (A) and WM (B), and for the LST toolbox in the bottom row for GM (C) and WM (D).

where the % of error in WM was up to 1.46% on images with >9 ml lesion load (see Fig. 4B). The error in WM increased with lesion volume on images where lesions were automatically segmented, but this was remarkably lower on the *SLS filled* images than those that were masked before segmentation (*SLS masked*). On the subset of images with >9 ml, the performance of the *SLS filled* was similar to that of the *Expert masked* pipeline.

Lesion volume also strongly correlated with the observed differences in normal-appearing GM and and WM for the *Original* ($r > 0.78$, $p < 0.001$) and *LST masked* ($r > 0.78$, $p < 0.001$) pipelines, and moderately for the *Expert masked* ($r > 0.36$, $p < 0.001$) and *LST filled* ($r > 0.40$, $p = 0.001$). As in the SLS, the error in GM and WM increased with lesion size on images where lesions were automatically segmented or intentionally left, and also increased remarkably in images where automatic lesion masks were masked instead of filled (see Fig. 4D). The % error in normal-appearing GM and WM of the *LST filled pipeline* was similar to that of the *Expert masked*.

3.3. Effect of lesion segmentation and filling

The lesion detection accuracy rate (true positives) of the SLS method was 0.43 ± 0.21, while the Dice similarity coefficient (Dice, 1945) between the estimated and manual annotated masks was 0.32 ± 0.17. The number of false positive lesion voxels (number of voxels misclassified as lesion), and false negative lesion voxels (number of missed lesion voxels) correlated with the % of error in total GM and WM volume of the *SLS filled* ($r > 0.60$, $p < 0.001$), and *LST filled* pipeline images ($r > 0.42$, $p < 0.001$). This suggested that in these pipelines, the observed error in tissue segmentation was mostly caused by the addition of false positive lesion voxels pertaining to GM that were filled with typical WM signal intensity, and also by the effect of missed hypo-intense WM lesion voxels into tissue distributions. In contrast, the % error in normal-appearing GM and WM in the images processed with the *SLS filled* and *LST filled* pipelines only correlated weakly with the number of false positives. Even some actual GM false positive voxels

were reassigned to WM, still WM voxels that were misclassified as lesion voxels were again reassigned to WM reducing the effect of false positives on the observed errors in normal-appearing tissue volume.

Similarly, the detection accuracy rate of the LST method was 0.41 ± 0.20, with Dice similarity coefficient of 0.35 ± 0.21. the number of false positives and false negatives correlated with the % of errors in total GM and WM of the *LST masked* pipeline ($r > 0.30$, $p = 0.01$), and only with the error in total GM of the *SLS masked* pipeline ($r > 0.52$, $p < 0.001$). Moreover, the number of false negative lesion voxels correlated weakly with the % of errors in GM and WM ($r > 0.40$, $p = 0.001$) of both pipelines. Contrary to filled images, actual GM voxels that were incorrectly classified as WM were not considered in tissue volume, reducing the linear correlation between the errors in lesion and tissue segmentation.

We also interchanged the lesion filling methods between the SLS and LST toolboxes and segmented again each set of images with the aim of evaluating the effect of each lesion filling process on the observed % differences in tissue volume. Differences were not statistically different with respect to the original pipelines for both GM and WM volume.

4. Discussion

The effect of lesions on total tissue volume was partly limited due to the canceling effect between the errors produced in normal-appearing tissue and the number of lesion voxels that were segmented as GM (Valverde et al., 2015b). This aspect is relevant because it explains why the observed % of error in total tissue volume was small or not significant between the evaluated pipelines of our study, even within the *Original* images intentionally segmented containing lesions. Furthermore, the % of error in total and normal-appearing WM volume in the images automatically segmented with either the SLS or LST was significantly lower when lesions voxels were filled than when they were masked before segmentation. As also reported in previous studies (Battaglini et al., 2012; Chard et al., 2010; Valverde et al., 2014), our results highlight the necessity to refill WM lesions before tissue segmentation for accurate cross-sectional tissue volume measurements.

However, the accuracy of automated lesion segmentation techniques is still low (Roura et al., 2015; Schmidt et al., 2012). Both automated pipelines overestimated normal-appearing WM (and underestimated GM) mostly by the effect of misclassified lesion voxels. Our results showed a significant but moderate correlation between underestimated total WM and the number of false positives of the *SLS filled* and *LST filled* pipelines. In contrast, the number of false positives correlated weakly with the differences in normal-appearing GM and WM, which might indicate that part of the false positive voxels that were actually WM were correctly reclassified after being filled. The % of error in the *SLS filled* and *LST filled* pipelines also correlated with the number of missed lesion voxels, which in addition to the clear correlation between the errors in tissue segmentation and lesion size, suggests that most of the differences observed in normal-appearing tissue volume were produced by the amount of missed lesion voxels that altered the tissue signal intensity distributions. This aspect suggests that the accuracy of new automatic tissue segmentation pipelines may be increased specially by reducing the number of missed lesion voxels, and in particular when those are hypo-intense in T1-w and should be filled before tissue segmentation. However, this study did not evaluate the methods with RRMS or SPSS image data, because the clinical focus of the study was on the initial CIS phenotype of MS, where paraclinical information is more relevant. In this regard, a further analysis of the accuracy of the evaluated pipelines on images with larger lesion load should be performed.

As expected, the *Expert masked* pipeline reported the lowest error in total and normal-appearing volume, although our results confirmed that masking out lesion voxels before tissue segmentation might not be optimal, as the error in tissue segmentation tends to increase with lesion size (Valverde et al., 2014). More interestingly, the performance of the fully automated *SLS filled* and *LST filled* pipelines was similar to that of the *Expert masked*, which seems to indicate that upon a certain lesion load, the errors produced by misclassified lesion voxels in the fully automated pipelines were comparable to the masking out error produced by not filling the expert annotations before tissue segmentation.

Within our data, the maximum differences in tissue volume produced by the *SLS filled* and *LST filled* might be lower than the own reproducibility of the SPM method, as stated in previous studies (De Boer et al., 2010; Nakamura et al., 2014). However, a direct comparison between studies has to be contemplated with care, because we did not perform a scan–reposition–rescan analysis of the evaluated pipelines, and consequently the differences in tissue volume produced by automated methods should be added to the inner reproducibility of the tissue segmentation method. Additionally, differences in the preprocessing pipelines between studies should be also contemplated, as shown in previous studies (Boyes et al., 2008; Zheng et al., 2009). The maximum differences in tissue volume produced by the fully automated pipelines also raises the question if the observed differences could be considered negligible when compared with the loss in tissue volume observed in follow-up scans. In this aspect, the differences in tissue volume shown by both the *SLS filled* and *LST filled* are remarkably lower than yearly tissue loss reported in recent clinical studies (Filippi et al., 2013; Sastre-Garriga et al., 2014; Uher et al., 2014). Hence, given the small error introduced by these methods, we recommend the use of either the SLS or LST toolkit.

There are a number of limitations in this work that have to be considered. This study was conducted using single-center data, and hence the applicability in a multi-center study was not determined here. The lack of manual tissue annotations does not allow us to analyze the tissue segmentation accuracy of each of the evaluated pipelines. Gold-standard annotations are time-consuming and have to be delineated by trained experts, a task which unfortunately is not always possible, especially when the number of subjects grows. In this aspect, the results of this study have to be understood under the premise that we are not evaluating the accuracy of the tissue segmentation methods, but the differences with respect to the manual expert pipeline that introduces the lowest error in tissue volume in images containing WM lesions (Battaglini et al., 2012; Valverde et al., 2014). The % of error in GM and WM volume introduced by the evaluated pipelines was small, and it was difficult to scale our findings with previous studies, given the differences in preprocessing and internal routines of each pipeline. Furthermore, in spite of the small error observed, our claims about the effectiveness of the fully automated pipelines have to be prudent, given the lesion load of the cohort of CIS patients of our study. As a future work, we will investigate the effect of images with higher lesion load on automated lesion segmentation, the posterior lesion filling process, and the impact of these automated processes in tissue segmentation methods.

In summary, this study shows that the automated lesion segmentation and filling methods included in the LST and SLS toolboxes reduce significantly the impact of T1-w hypo-intense lesions on the SPM8 tissue segmentation method. Our results show that compared with the evaluated pipelines that require manual expert intervention, the accuracy in tissue segmentation is not affected remarkably on images processed with the fully automated pipelines. This is relevant and suggests that LST and SLS toolboxes allow for performing accurate brain tissue volume measurements without any kind of manual intervention. The possibility of filling MS white matter lesions without manual delineation of lesions is pertinent not only in terms of time and economic costs, but also to avoid the inherent *intra*/*inter* variability between manual annotations.

Acknowledgments

S. Valverde holds a FI-DGR2013 grant from the Generalitat de Catalunya. E. Roura holds a BR-UdG2013 grant. This work has been partially

supported by “La Fundació la Marató de TV3” and by Retos de Investigación TIN2014-55710-R.

References

Ashburner, J., Friston, K., 2005. Unified segmentation. NeuroImage 26, 839–851.

Battaglini, M., Jenkinson, M., De Stefano, N., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33 (9), 2062–2071.

Boyes, R., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D., Bernstein, M.A., et al., 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage 39, 1752–1762.

Cabezas, M., Oliver, A., Roura, E., Freixenet, J., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2014. Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding. Comput. Methods Prog. Biomed. 115 (3), 147–161.

Ceccarelli, A., Jackson, J.S., Tauhid, S., Arora, A., Gorky, J., Dell'Oglio, E., Bakshi, A., et al., 2012. The impact of lesion in-painting and registration methods on voxel-based morphometry in detecting regional cerebral gray matter atrophy in multiple sclerosis. Am. J. Neuroradiol. 33 (8), 1579–1585.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32, 223–228.

Clark, K.A., Woods, R.P., Rottenberg, D.A., Toga, A.W., Mazziotta, J.C., 2006. Impact of acquisition protocols and processing streams on tissue segmentation of T1-w weighted MR images. NeuroImage 29, 185–202.

De Boer, R., Vrooman, H., Ikram, M., Vernooij, M., Breteler, M., Van der Lugt, A., Niessen, W., 2010. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. NeuroImage 51 (3), 1047–1056.

De Bresser, J., Portegies, M.P., Leemans, A., Biessels, G.J., Kappelle, L.J., Viergever, M.A., 2011. A comparison of MR based segmentation methods for measuring brain atrophy progression. NeuroImage 54 (2), 760–768.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Filippi, M., Preziosa, P., Copetti, M., Riccitelli, G., Horsfield, M.A., Martinelli, V., Comi, G., Rocca, M.A., 2013. Gray matter damage predicts the accumulation of disability 13 years later in MS. Neurology 81 (20), 1759–1967.

Ganiler, O., Oliver, A., Díez, Y., Freixenet, J., Vilanova, J.C., Beltrán, B., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2014. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. Neuroradiology 56 (5), 363–374.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17 (1), 1–18.

Guizard, N., Coupé, P., Fonov, V., Manjón, J., Arnold, D.L., Collins, D.L., 2015. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. NeuroImage Clin. 8, 376–389.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, A., 2012. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Inf. Sci. 186 (1), 164–185.

Nakamura, K., Fisher, E., 2009. Segmentation of brain magnetic resonance images for measurement of gray matter atrophy in multiple sclerosis patients. NeuroImage 44 (3), 769–776.

Nakamura, K., Guizard, N., Fonov, V.S., Narayanan, S., Collins, D.L., Arnold, D.L., 2014. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. NeuroImage Clin. 4, 10–17.

Pérez-Miralles, F., Sastre-Garriga, J., Tintoré, M., Arrambide, G., Nos, C., Perkal, H., Río, J., et al., 2013. Clinical impact of early brain atrophy in clinically isolated syndromes. Mult. Scler. 19 (14), 1878–1886.

Popescu, V., Battaglini, M., Hoogstrate, W.S., et al., 2012. Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. NeuroImage 61 (4), 1484–1494.

Popescu, V., Ran, N.C.G., Barkhof, F., Chard, D.T., Wheeler-Kingshott, C.A., Vrenken, H., 2014. Accurate GM atrophy quantification in MS using lesion filling with co-registered 2D lesion masks. NeuroImage Clin. 4 (January), 366–373.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology http://dx.doi.org/10.1007/s00234-015-1552-2.

Sanfilipo, M.P., Benedict, R.B., Sharma, J., Weinstock-Guttman, B., Bakshi, R., 2005. The relationship between whole brain volume and disability in multiple sclerosis: a comparison of normalized gray vs. white matter with misclassification correction. NeuroImage 26 (4), 1068–1077.

Sastre-Garriga, J., Tur, C., Pareto, D., Vidal-Jordana, A., Auger, C., Río, J., Huerga, E., Tintoré, M., Rovira, A., Montalban, X., 2014. Brain atrophy in natalizumab-treated patients: a 3-year follow-up. Mult. Scler. http://dx.doi.org/10.1177/1352458514556300.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., et al., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. NeuroImage 59 (4), 3774–3783.

Sled, J.G., Zijdenbos, P., Evans, C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17 (1), 479–489.

Uher, T., Horakova, D., Bergsland, N., Tyblova, M., Ramasamy, D.P., Seidl, Z., Vaneckova, M., Krasensky, J., Havrdova, E., Zivadinov, R., 2014. MRI correlates of disability progression in patients with CIS over 48 months. NeuroImage: Clinical 6 (January). Elsevier B.V., pp. 312–319.

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion filling approach to improve brain tissue volume measurements. NeuroImage Clin. 6 (January), 86–92.

Valverde, S., Oliver, A., Cabezas, M., Roura, E., Lladó, X., 2015a. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. J. Magn. Reson. Imaging 41 (1), 93–101.

Valverde, S., Oliver, A., Díez, Y., Cabezas, M., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015b. Evaluating the effects of white matter multiple sclerosis lesions on the volume estimation of 6 brain tissue segmentation methods. Am. J. Neuroradiol. http://dx.doi.org/10.3174/ajnr.A4262.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.

Zheng, W., Chee, M.W., Zagorodnov, V., 2009. Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. NeuroImage 48, 73–83.